Biometrika (1973), **60**, 3, *p.* 607
Printed in Great Britain

Optimum Monte-Carlo sampling using Markov chains

BY P. H. PESKUN
York University, Toronto

SUMMARY

The sampling method proposed by Metropolis *et al.* (1953) requires the simulation of a Markov chain with a specified $\boldsymbol{\pi}$ as its stationary distribution. Hastings (1970) outlined a general procedure for constructing and simulating such a Markov chain. The matrix $\mathbf{P}$ of transition probabilities is constructed using a defined symmetric function s_{ij} and an arbitrary transition matrix $\mathbf{Q}$. Here, for a given $\mathbf{Q}$, the relative merits of the two simple choices for s_{ij} suggested by Hastings (1970) are discussed. The optimum choice for s_{ij} is shown to be one of these. For the other choice, those $\mathbf{Q}$ are given which are known to make the sampling method based on $\mathbf{P}$ asymptotically less precise than independent sampling.

Some key words: Monte-Carlo estimation; Markov chain method of sampling; Variance reduction; Simulation.

1. INTRODUCTION

Suppose we wish to estimate the expectation

$$I = E_\pi(f) = \sum_{i=0}^{S} f(i)\pi_i,$$

where $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_S)$ is a positive probability distribution, i.e. $\pi_i > 0$ for all i, and $f(\cdot)$ is a nonconstant function defined on the states $0, 1, \ldots, S$ of an irreducible Markov chain determined by the transition matrix $\mathbf{P} = \{p_{ij}\}$. Throughout this paper, $\boldsymbol{\pi}$ and $f(\cdot)$ are to be considered fixed.

If $\mathbf{P}$ is chosen so that $\boldsymbol{\pi}$ is its unique stationary distribution, i.e. $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, then after simulating the Markov chain for times $t = 1, \ldots, N$, an estimate of the expectation I is given by

$$\hat{I} = \sum_{t=1}^{N} f\{X(t)\}/N,$$

where $X(t)$ denotes the state occupied by the chain at time t.

Hastings (1970, p. 99) outlines a general procedure for constructing such a transition matrix $\mathbf{P} = \{p_{ij}\}$. First of all, $\mathbf{P}$ is required to satisfy, for all pairs of states i and j, the reversibility condition

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{1}$$

Secondly, it is assumed that p_{ij} has the form

$$p_{ij} = q_{ij}\alpha_{ij} \quad (i \neq j), \tag{2}$$

with

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij},$$

where $\mathbf{Q} = \{q_{ij}\}$ is the transition matrix of an arbitrary irreducible Markov chain on the states $0, 1, \ldots, S$. Also, α_{ij} is given by

$$\alpha_{ij} = \frac{s_{ij}}{1 + t_{ij}}, \tag{3}$$

where s_{ij} is a symmetric function of i and j chosen so that $0 \leqslant \alpha_{ij} \leqslant 1$ for all i and j, and $t_{ij} = (\pi_i q_{ij})/(\pi_j q_{ji})$.

The Markov chain determined by a transition matrix $\mathbf{P}$ of the above form is simulated by carrying out the following steps for each time t:

(i) assume that $X(t) = i$ and select a state j using the distribution given by the ith row of $\mathbf{Q}$;

(ii) take $X(t+1) = j$ with probability α_{ij} and $X(t+1) = i$ with probability $1-\alpha_{ij}$.

From (2) and (3) we see that the symmetric function s_{ij} and the arbitrary transition matrix $\mathbf{Q}$ determine the transition matrix $\mathbf{P}$. The purpose of this paper is to determine the optimum choice of s_{ij}, for a given choice of $\mathbf{Q}$, so that the estimate $\hat{I}$ is as precise as possible. The relative merits of the two simple choices $s_{ij}^{(M)}$ and $s_{ij}^{(B)}$ for s_{ij}, suggested by Hastings (1970, p. 100), are discussed. For a given choice of $\mathbf{Q}$, the sampling method based on $\mathbf{P}$ is shown to be asymptotically as precise as possible for $s_{ij} = s_{ij}^{(M)}$. For $s_{ij} = s_{ij}^{(B)}$, those $\mathbf{Q}$ are given which are known to make the sampling method based on $\mathbf{P}$ asymptotically less precise than independent sampling.

2. The symmetric function s_{ij}

2·1. *The asymptotic variance and bias of the estimate* $\hat{I}$

The following definition is due to Kemeny & Snell (1969, p. 75).

Definition 2·1·1. *The matrix* $\mathbf{A} = \boldsymbol{\xi}^T\boldsymbol{\pi}$, *where* $\boldsymbol{\xi} = (1, 1, \ldots, 1)$, *and the inverse matrix* $\mathbf{Z} = \{\mathbf{I} - (\mathbf{P} - \mathbf{A})\}^{-1}$ *will be called the 'limiting matrix' and the 'fundamental matrix', respectively, for the finite irreducible Markov chain determined by the transition matrix* $\mathbf{P}$ *whose stationary distribution is* $\boldsymbol{\pi}$.

Let $\mathbf{f}$ be the $1 \times (S+1)$ row vector $\mathbf{f} = \{f(0), f(1), \ldots, f(S)\}$ and let $\mathbf{B} = \{b_{ij}\}$ be the $(S+1) \times (S+1)$ diagonal matrix with diagonal vector $\boldsymbol{\pi}$; that is, $b_{ii} = \pi_i$ $(i = 0, 1, \ldots, S)$.

Even though the variance and bias of the estimate $\hat{I} = [f\{X(1)\} + \ldots + f\{X(N)\}]/N$ cannot be expressed as functions of the sample size N which lend themselves to easy analysis, asymptotic expressions for these two quantities can be obtained which will be useful not only analytically but also practically since, in actual simulations, the sample size N will usually be large.

The following asymptotic expression for the variance of the estimate $\hat{I} = \Sigma f\{X(t)\}/N$ which Kemeny & Snell (1969, p. 84) have derived is independent of the distribution $\boldsymbol{\gamma}$ of the initial state $X(0)$:

$$\begin{aligned} v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}) &= \lim_{N\to\infty} N \operatorname{var}\left[\sum_{t=1}^{N} f\{X(t)\}/N\right] \\ &= \mathbf{f}\{\mathbf{BZ} + (\mathbf{BZ})^T - \mathbf{B} - \mathbf{BA}\}\mathbf{f}^T \\ &= \mathbf{f}(2\mathbf{BZ} - \mathbf{B} - \mathbf{BA})\,\mathbf{f}^T. \end{aligned} \tag{4}$$

In what follows, we shall assess the precision of the estimate $\hat{I}$ by $v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})$, it being assumed that the sample size N is sufficiently large so that the error in the approximation

$$\operatorname{var}(\hat{I}) \doteq v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})/N$$

is small. We shall refer to $v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})$ as the asymptotic variance of the estimate $\hat{I}$.

With respect to the bias of the estimate $\hat{I}$, it can be shown that, as $N \to \infty$, $\lim N\{E(\hat{I}) - I\}$ exists and is dependent on the distribution $\boldsymbol{\gamma}$ of the initial state $X(0)$. Since var $(\hat{I})$ is $O(N^{-1})$ and the bias squared is $O(N^{-2})$, we feel that, for an appropriate distribution $\boldsymbol{\gamma}$ of the initial state $X(0)$, the bias has a negligible effect on the accuracy of the estimate $\hat{I}$ and is not an appreciable disadvantage of the Markov chain method of sampling; in fact, if $X(0)$ is sampled from $\boldsymbol{\pi}$ itself, then $\hat{I}$ is unbiased. We shall thus confine our discussions to the precision, rather than the accuracy, of the estimate $\hat{I}$.

In the following theorem, which gives a sufficient condition for asymptotic variance reduction and which will be useful in determining the optimum symmetric function s_{ij}, we define $\mathbf{P}_2 \leqslant \mathbf{P}_1$ if each of the off-diagonal elements of $\mathbf{P}_1$ is greater than or equal to the corresponding off-diagonal elements of $\mathbf{P}_2$.

THEOREM 2·1·1. *Suppose each of the irreducible transition matrices* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *satisfies the reversibility condition* (1) *for the same probability distribution* $\boldsymbol{\pi}$. *If* $\mathbf{P}_2 \leqslant \mathbf{P}_1$, *then for the estimate* $\hat{I} = \sum_{t=1}^{N} f\{X(t)\}/N$,

$$v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}_1) \leqslant v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}_2).$$

Proof. For the (k, l)th off-diagonal element p_{kl} of $\mathbf{P}$, we have from (4),

$$\frac{\partial\{v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})\}}{\partial p_{kl}} = 2\mathbf{f}\left(\mathbf{B}\frac{\partial \mathbf{Z}}{\partial p_{kl}}\right)\mathbf{f}^T \quad (k \neq l).$$

Since $\mathbf{P}$ satisfies the reversibility condition (1), it follows that the matrix $\mathbf{BP}$ is symmetric. Similarly, the matrix $\mathbf{BA}$ is symmetric. The symmetry of the matrices

$$\mathbf{BZ}^{-1} = \mathbf{B} - \mathbf{BP} + \mathbf{BA}, \quad \mathbf{ZB}^{-1} = (\mathbf{BZ}^{-1})^{-1}$$

implies the symmetry of the matrix $\mathbf{BZ} = \mathbf{B}(\mathbf{ZB}^{-1})\,\mathbf{B}$. If we substitute

$$\frac{\partial \mathbf{Z}}{\partial p_{kl}} = -\mathbf{Z}\frac{\partial \mathbf{Z}^{-1}}{\partial p_{kl}}\mathbf{Z},$$

and use the symmetry of $\mathbf{BZ}$ and $\mathbf{B}$, we then have

$$\frac{\partial\{v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})\}}{\partial p_{kl}} = -2(\mathbf{Zf}^T)^T\left(\mathbf{B}\frac{\partial \mathbf{Z}^{-1}}{\partial p_{kl}}\right)(\mathbf{Zf}^T) \quad (k \neq l).$$

Since $\mathbf{BZ}^{-1} = \mathbf{B} - \mathbf{BP} + \mathbf{BA}$, $\mathbf{P}$ being a transition matrix satisfying the reversibility condition (1), then all the elements of the matrix $\mathbf{B}(\partial \mathbf{Z}^{-1}/\partial p_{kl}) = -\mathbf{B}(\partial \mathbf{P}/\partial p_{kl})$ $(k \neq l)$ are equal to zero except for the (l, l), (l, k), (k, l) and (k, k)th elements, which are equal to π_k, $-\pi_k$, $-\pi_k$ and π_k, respectively. It follows that the matrix $\mathbf{B}(\partial \mathbf{Z}^{-1}/\partial p_{kl})$ $(k \neq l)$ is positive semi-definite with one nonzero eigenvalue equal to $2\pi_k$. Thus we have

$$\frac{\partial\{v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})\}}{\partial p_{kl}} = -2(\mathbf{Zf}^T)^T\left(\mathbf{B}\frac{\partial \mathbf{Z}^{-1}}{\partial p_{kl}}\right)(\mathbf{Zf}^T) \leqslant 0 \quad (k \neq l).$$

This result implies that the asymptotic variance $v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})$ is a decreasing function in the off-diagonal elements of $\mathbf{P}$ and thus it follows that $\mathbf{P}_2 \leqslant \mathbf{P}_1$ implies $v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}_1) \leqslant v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}_2)$.

For a large sample size N, Theorem 2·1·1 suggests that the variance of the estimate $\hat{I}$ can be reduced by appropriately transferring weight from the diagonal elements of $\mathbf{P}$ to the off-diagonal elements. Intuitively, this makes sense. If the diagonal elements of $\mathbf{P}$ are small, then the probability of remaining in any given state will be small. This suggests an improvement in the sampling of all possible states which, in turn, suggests an improvement in the precision of the estimate $\hat{I}$.

2·2. *The optimum symmetric function s_{ij}*

Hastings (1970, p. 100) suggests two simple choices for the symmetric function s_{ij} which, for all i and j, are given by

(i) $s_{ij}^{(M)} = \begin{cases} 1+t_{ij} & (t_{ji} \geqslant 1), \\ 1+t_{ji} & (t_{ji} \leqslant 1), \end{cases}$

(ii) $s_{ij}^{(B)} = 1$.

For a symmetric $\mathbf{Q}$, $s_{ij} = s_{ij}^{(M)}$ gives the sampling method devised by Metropolis *et al.* (1953) and $s_{ij} = s_{ij}^{(B)}$ gives Barker's (1965) sampling method.

DEFINITION 2·2·1. *Let $\mathbf{P}^{(M)} = \{p_{ij}^{(M)}\}$ and $\mathbf{P}^{(B)} = \{p_{ij}^{(B)}\}$ denote the irreducible transition matrices constructed according to* (2) *and* (3) *using the same transition matrix $\mathbf{Q} = \{q_{ij}\}$ and the symmetric functions $s_{ij}^{(M)}$ and $s_{ij}^{(B)}$, respectively.*

From (3) we have

$$s_{ij} \leqslant 1 + \min(t_{ij}, t_{ji}), \tag{5}$$

since $\alpha_{ij} \leqslant 1$ implies $s_{ij} \leqslant 1+t_{ij}$ and $\alpha_{ji} \leqslant 1$ implies $s_{ij} = s_{ji} \leqslant 1+t_{ji}$. We note that equality is attained in (5) for the symmetric function $s_{ij} = s_{ij}^{(M)}$; that is,

$$s_{ij}^{(M)} = 1 + \min(t_{ij}, t_{ji}). \tag{6}$$

THEOREM 2·2·1. *For a given transition matrix $\mathbf{Q} = \{q_{ij}\}$, the optimum symmetric function s_{ij} is $s_{ij}^{(M)}$.*

Proof. For a given transition matrix $\mathbf{Q} = \{q_{ij}\}$, we have seen how to construct a transition matrix $\mathbf{P} = \{p_{ij}\}$, where $p_{ij} = q_{ij}\alpha_{ij}$ $(i \neq j)$ and $\alpha_{ij} = s_{ij}/(1+t_{ij})$. From (5) and (6) it follows that the maximum value for α_{ij} occurs for $s_{ij} = s_{ij}^{(M)}$. Thus, it is clear that $\mathbf{P} \leqslant \mathbf{P}^{(M)}$ for any irreducible transition matrix $\mathbf{P}$ where both $\mathbf{P}^{(M)}$ and $\mathbf{P}$ are constructed according to (2) and (3) using the same given transition matrix $\mathbf{Q}$. By Theorem 2·1·1 it follows that the optimum symmetric function s_{ij} is $s_{ij}^{(M)}$ since $v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}^{(M)}) \leqslant v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P})$.

2·3. *A comparison of the sampling methods based on $\mathbf{P}^{(M)}$ and $\mathbf{P}^{(B)}$*

From Theorem 2·2·1 and the definitions of the symmetric functions $s_{ij}^{(M)}$ and $s_{ij}^{(B)}$, it can be shown that, for a given transition matrix $\mathbf{Q}$,

$$v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}^{(M)}) \leqslant v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}^{(B)}) \leqslant v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{A}) + 2v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}^{(M)}),$$

where $v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{A})$ is the asymptotic variance for independent sampling, i.e. the theoretical independent sampling variance for sample size N equal to 1. For the special case where the given transition matrix $\mathbf{Q}$ itself satisfies the reversibility condition, i.e. $\pi_i q_{ij} = \pi_j q_{ji}$ for all i and j, it can then be shown that

$$v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}^{(B)}) = v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{A}) + 2v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}^{(M)}). \tag{7}$$

We note that independent sampling is just a special case of the sampling method based on $\mathbf{P}^{(M)}$ since $\mathbf{P}^{(M)} = \mathbf{A}$ for $\mathbf{Q} = \mathbf{A}$. This suggests the possibility of the sampling method based on $\mathbf{P}^{(M)}$, for an appropriate choice of $\mathbf{Q}$, being asymptotically more precise than independent sampling. If such is the case and if, in addition, $\mathbf{Q}$ itself satisfies the reversibility condition, then from (7) we see that asymptotically no matter how much more precise the sampling method based on $\mathbf{P}^{(M)}$ is than independent sampling, the sampling method based on $\mathbf{P}^{(B)}$ can

at best be only as precise as independent sampling. We shall now show that this is also true for the case where $\mathbf{Q}$ is symmetric, i.e. for Barker's (1965) sampling method.

In general, the sampling method based on $\mathbf{P}$, where $\mathbf{P}$ satisfies the reversibility condition (1), will be asymptotically as precise or more precise than independent sampling if and only if

$$\mathbf{fBZf}^T \leqslant \mathbf{fBf}^T, \tag{8}$$

where $\mathbf{Z}$ is the fundamental matrix determined by $\mathbf{P}$. Rewriting (4) in the following form, where $\mathbf{BA} = \boldsymbol{\pi}^T\boldsymbol{\pi}$,

$$\mathbf{fBZf}^T = \tfrac{1}{2}\{v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}) + \mathbf{fBf}^T + (\boldsymbol{\pi}\mathbf{f}^T)^T(\boldsymbol{\pi}\mathbf{f}^T)\},$$

we see that the symmetric matrix $\mathbf{BZ}$ is positive definite since the first and third terms of the above equality are nonnegative and the diagonal matrix $\mathbf{B}$ is positive definite. We refer the reader to Gantmacher (1960, p. 310) for the theory of pencils of quadratic forms which we will use in order to compare the quadratic forms $\mathbf{fBZf}^T$ and $\mathbf{fBf}^T$.

If we number the characteristic values of the regular pencil of forms $\mathbf{fBf}^T - \lambda\mathbf{fBZf}^T$ in nondecreasing order, then for $\lambda_0 \leqslant \lambda_1 \leqslant \ldots \leqslant \lambda_S$ it follows that for all functions $f(\cdot)$,

$$\lambda_0 \leqslant \mathbf{fBf}^T/\mathbf{fBZf}^T \leqslant \lambda_S.$$

Thus, we see that (8) is satisfied if and only if $\lambda_0 \geqslant 1$.

The characteristic equation of the regular pencil of forms $\mathbf{fBf}^T - \lambda\mathbf{fBZf}^T$ is $|\mathbf{B} - \lambda\mathbf{BZ}| = 0$, which can be written as

$$|\mu\mathbf{I} - (\mathbf{P} - \mathbf{A})| = 0,$$

where $\mu = 1 - \lambda$. It thus follows that the matrix $\mathbf{P} - \mathbf{A}$ has $S+1$ real characteristic roots, since a regular pencil of forms always has real roots. Since $\mathbf{P}$ and $\mathbf{A}$ are transition matrices, we have

$$(\mathbf{P} - \mathbf{A})\boldsymbol{\xi}^T = \mathbf{P}\boldsymbol{\xi}^T - \mathbf{A}\boldsymbol{\xi}^T = \boldsymbol{\xi}^T - \boldsymbol{\xi}^T = 0\boldsymbol{\xi}^T.$$

Hence, $\mu = 0$ is a characteristic root of the matrix $\mathbf{P} - \mathbf{A}$. Also, since $\mu \leqslant 0$ implies $\lambda \geqslant 1$, we then have proved the following theorem.

THEOREM 2·3·1. *For any function $f(\cdot)$, the Markov chain sampling method based on $\mathbf{P}$ will be asymptotically as precise or more precise than independent sampling with respect to the positive distribution $\boldsymbol{\pi}$, i.e.*

$$v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{P}) \leqslant v(\mathbf{f}, \boldsymbol{\pi}, \mathbf{A}),$$

if and only if the nonzero characteristic roots of the matrix $(\mathbf{P} - \mathbf{A})$ are negative.

Since the sum of the diagonal elements of the matrix $\mathbf{P}^{(B)}$ is equal to the sum of all its elements minus the sum of its off-diagonal elements, i.e.

$$S + 1 - \sum\sum_{i>j} \{p_{ij}^{(B)} + p_{ji}^{(B)}\},$$

and since the sum of the diagonal elements of the matrix $\mathbf{A}$ is 1, then the sum of the characteristic roots of the matrix $\mathbf{P}^{(B)} - \mathbf{A}$ is

$$S - \sum_{i>j} \{p_{ij}^{(B)} + p_{ji}^{(B)}\}.$$

For a given $\mathbf{Q} = \{q_{ij}\}$, it can be shown that for $i \neq j$

$$\min(q_{ij}, q_{ji}) \leqslant p_{ij}^{(B)} + p_{ji}^{(B)} \leqslant \max(q_{ij}, q_{ji}),$$

with both inequalities becoming equalities if $\mathbf{Q}$ is symmetric. Hence, it follows that if $\mathbf{Q}$ is symmetric then

$$S - \sum_{i>j} \{p_{ij}^{(B)} + p_{ji}^{(B)}\} = S - \sum_{i>j} q_{ij} \geqslant \tfrac{1}{2}(S-1), \tag{9}$$

since the sum of the lower off-diagonal elements of $\mathbf{Q}$ can be at most $\frac{1}{2}(S+1)$.

Except for the two-state system, i.e. $S = 1$, we see from (9) that the sum of the characteristic roots of the matrix $\mathbf{P}^{(B)} - \mathbf{A}$ is positive. This implies that there must be at least one positive characteristic root among its nonzero characteristic roots. For $S > 1$, we have thus shown that Barker's (1965) sampling method is asymptotically less precise than independent sampling. For $S = 1$, it can be shown that Barker's (1965) sampling method is asymptotically, at best, as precise as independent sampling if $\mathbf{Q}$ is chosen so that $q_{01} = q_{10} = 1$.

This work started in my Ph.D. thesis, University of Toronto, and was completed with the support of a grant from the National Research Council of Canada. I would like to thank my thesis supervisor, Professor W. K. Hastings, for his continuous guidance and many useful suggestions. Thanks are also due to the referees whose comments were very helpful.

REFERENCES

BARKER, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Austral. J. Phys.* **18**, 119–33.

GANTMACHER, F. R. (1960). *The Theory of Matrices*, vol. I. New York: Chelsea.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

KEMENY, J. G. & SNELL, J. L. (1969). *Finite Markov Chains*. Princeton: Van Nostrand.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–92.

[*Received May* 1972. *Revised May* 1973]